DOCUMENT RESUME

ED 435 691 TM 030 322

AUTHOR van Krimpen-Stoop, Edith M. L. A.; Meijer, Rob R.
TITLE CUSUM-Based Person-Fit Statistics for Adaptive Testing. Research Report 99-05.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
SPONS AGENCY Law School Admissions Council, Newtown, PA.
PUB DATE 1999-00-00
NOTE 30p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Estimation (Mathematics); *Item Response Theory
IDENTIFIERS *Person Fit Measures

ABSTRACT

Item scores that do not fit an assumed item response theory model may cause the latent trait value to be estimated inaccurately. Several person-fit statistics for detecting nonfitting score patterns for paper-and-pencil tests have been proposed. In the context of computerized adaptive tests (CAT), the use of person-fit analysis has hardly been explored. In this study, new person-fit statistics are proposed, and critical values for these statistics are derived from existing statistical theory. Statistics are proposed that are sensitive to runs of correct or incorrect item scores and are based on all items administered in a CAT or based on subsets of items, using observed and expected item scores and using cumulative sum (CUSUM) procedures. The theoretical and empirical distributions of the statistics are compared and detection rates are investigated. Results show that the nominal and empirical Type I error rates are comparable for CUSUM procedures when the number of items in each subset and the number of measurement points are not too small. Detection rates of CUSUM procedures were superior to other fit statistics. Applications of the statistics are discussed. (Contains 5 tables and 21 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 435 691

CUSUM-Based Person-Fit Statistics for Adaptive Testing

Research Report 99-05

Edith M.L.A. van Krimpen-Stoop
Rob R. Meijer

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Helissen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030322

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

BEST COPY AVAILABLE

CUSUM-Based Person-Fit Statistics
for Adaptive Testing

Edith M.L.A. van Krimpen-Stoop
Rob R. Meijer

Abstract

Item scores that do not fit an assumed item response theory model may cause the latent trait value to be inaccurately estimated. Several person-fit statistics for detecting nonfitting score patterns for paper-and-pencil tests have been proposed. In the context of computerized adaptive tests (CAT), the use of person-fit analysis is hardly explored. Because it has been shown that the distribution of existing person-fit statistics is not applicable in a CAT, in this study new person-fit statistics are proposed and critical values for these statistics are derived from existing statistical theory. Statistics are proposed that are sensitive to runs of correct or incorrect item scores and are based on all items administered in a CAT or based on subsets of items, using observed and expected item scores and using cumulative sum (CUSUM) procedures. The theoretical and empirical distributions of the statistics are compared and detection rates are investigated. Results showed that the nominal and empirical type I error rates were comparable for CUSUM procedures when the number of items in each subset and the number of measurement points were not too small. Detection rates of CUSUM procedures were superior to other fit statistics. Applications of the statistics are discussed.

Key words: appropriateness measurement, computer adaptive testing, cumulative sum, item response theory, person fit.

CUSUM-Based Person-Fit Statistics for Adaptive Testing

An examinee's test score may not reveal the operation of undesirable influences of test taking behavior such as faking on biodata questionnaires and personality tests, or guessing, or knowledge of the correct answers due to test preview on achievement tests. These and other influences may result in inappropriate test scores which may have serious consequences for practical test use, for example, in job and educational selection, where classification errors may result.

In the context of item response theory (IRT) modeling, several methods have been proposed to detect item score patterns that are not in agreement with the expected item score pattern based on a particular test model. These item score patterns should be detected because scores of such persons may not be adequate descriptions of their trait level (θ). This area of research is commonly referred to as person-fit research, and most person-fit studies have concentrated on the development of fit statistics that can be used to identify nonfitting response vectors; examples of fit statistics can be found in Levine and Rubin (1979), Drasgow, Levine, and Williams (1985), Levine & Drasgow (1988), and Molenaar and Hoijtink (1990).

Most fit statistics have been proposed in the context of paper-and-pencil (P&P) tests. In almost all statistics, for an individual examinee with a latent trait value θ, the residual of the observed and expected item scores on the basis of an IRT model is compared across items. When, under the null model of fitting response behavior, the null distribution of a statistic is known, item score patterns can be classified as fitting or nonfitting.

Recently, the use of person-fit statistics has been explored in the context of adaptive testing. Nering (1997) and van Krimpen-Stoop and Meijer (in press-a) showed that the use of existing person-fit statistics in adaptive testing is not straightforward. One of the problems is, that the characteristics of a computerized adaptive test (CAT) are unfavorable for the assessment of person fit (Reise & Due, 1991, Molenaar & Hoijtink, 1990, 1996). A CAT contains relatively few items compared with a P&P test. Because the detection rate is sensitive to test length - longer tests will result in higher detection rates (e.g., Meijer, Molenaar, & Sijtsma, 1994) - the detection rate for a CAT will, in general be lower than for a P&P test. A second problem is that almost all person-fit statistics use the spread of the item difficulties: generally speaking, nonfitting item score patterns consists of many

incorrect (0) scores to easy items and many correct (1) scores to difficult items. In a CAT, the spread in the item difficulties is relatively modest: in particular at the end of the test when the estimate of θ, $\hat{\theta}$, is close to true θ, items with similar item difficulties will be selected and as a result it will be difficult to distinguish fitting from nonfitting item score patterns. Due to the modest spread in the item difficulties, the assumed null distribution of most existing person-fit statistics is also underdispersed, which results in empirical type I errors that are too small compared to the nominal type I errors (van Krimpen-Stoop and Meijer, in press-a). Furthermore, for both P&P testing and adaptive testing, the null distribution of existing person-fit statistics varies across different θ-values (Reise, 1995, van Krimpen-Stoop & Meijer, in press-a). As a result, it is difficult to use one critical value for all examinees to classify item score patterns as nonfitting.

In this study, statistics will be proposed that are especially designed for use in a CAT and critical values for these statistics will be derived on the basis of which item score patterns can be classified as fitting or nonfitting. This paper is organized as follows. First, existing literature on person-fit statistics in adaptive testing is discussed. Second, new statistics are proposed. Third, the distribution of these statistics is obtained from existing statistical theory and by means of simulation studies, the nominal type I errors are compared with the empirical type I errors. Finally, the detection rates of these statistics are investigated.

Item Response Theory and Person-Fit Research

IRT models describe the probability of a correct response to an item as a function of item and person parameters (e.g., Hambleton & Swaminathan, 1985, pp. 35-48). Let X_i be the binary $(0,1)$ response to item i $(i = 1, ..., I)$, where 1 denotes a correct or keyed response, and 0 denotes an incorrect or not keyed response. Further, let a_i denote the item discrimination parameter, b_i the item difficulty parameter, and θ the latent trait value. The probability of correctly answering an item according to the two-parameter logistic IRT model (2-PLM) can be written as

$$P_i(X_i = 1 \,|\theta, a_i, b_i) = P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}. \quad (1)$$

In this study we use the 2-PLM because it is less restrictive with respect to empirical data

than the one-parameter logistic model and it does not have the estimation problems of the guessing parameter in the three-parameter logistic model (e.g., Baker, 1992, pp.109-112). The results obtained in this study, however, are easily generalized to the one- and three-parameter logistic model.

In a P&P test the same items are administered to each examinee. To investigate an examinee's fit to an IRT model in almost all person-fit statistics the residual of the observed and expected item score is compared across items. Let $w_i(\theta)$ and $w_0(\theta)$ be suitable functions. Following Snijders (in press), a general form in which most person-fit statistics can be expressed is

$$\sum_{i=1}^{I} X_i w_i(\theta) - w_0(\theta). \tag{2}$$

To have a person-fit statistic with expectation 0, many person-fit statistics are expressed in the centered form

$$W(\theta) = \sum_{i=1}^{I} [X_i - P_i(\theta)] w_i(\theta) \tag{3}$$

which results in a function of the residuals of the observed and expected item scores. Note that, as a result of binary scoring $X_i^2 = X_i$; thus, for a suitable function $v_i(\theta)$, statistics of the form

$$\sum_{i=1}^{I} [X_i - P_i(\theta)]^2 v_i(\theta),$$

can be re-expressed as statistics of the form in Equation 2. Often the variance of the statistic is taken into account to obtain a standardized version of the statistic. For example, Wright and Stone (1979) proposed person-fit statistics based on squared standardized residuals, where the residuals are weighted with the variances of the item scores of the I items

$$V(\theta) = \sum_{i=1}^{I} \frac{[X_i - P_i(\theta)]^2}{P_i(\theta) Q_i(\theta)},$$

where $Q_i(\theta) = 1 - P_i(\theta)$.

For classifying an item score pattern as nonfitting, an important tool is the probability

of exceedance or significance probability. Let t be the observed value of the person-fit statistic T and assume that larger values of T denote nonfitting item score patterns. Then, the one-sided significance probability is defined as the probability that the value of the test statistic is larger than the observed value: $p^* = P(T \geq t)$. When $p^* < \alpha$, where α is the one-sided significance level, an item score pattern can be classified as nonfitting. The value of a test statistic with $p^* = \alpha$ will be denoted as the critical value at one-sided significance level α.

Person-Fit Research in CAT

Van Krimpen-Stoop and Meijer (in press-a) (see also Nering, 1997) investigated the empirical distribution of an often used fit statistic in the context of P&P tests, the standardized log-likelihood statistic l_z (Drasgow, Levine, & Williams, 1985) and an adaptation l_z^* (Snijders, in press) that corrects for the use of the estimated $\hat{\theta}$ instead of true θ in l_z. Both statistics were assumed to be asymptotically standard normally distributed. Van Krimpen-Stoop and Meijer (in press-a) found that, for simulated P&P data when $\hat{\theta}$ was used instead of θ, the nominal and empirical type I errors of l_z^* were similar, whereas these errors for l_z were different, especially for short tests. For CAT data, however, there was a large discrepancy between the empirical and theoretical distribution for both statistics. Consequently, decisions about the fit of a score pattern on the basis of theoretical critical values were inappropriate. As an alternative, van Krimpen-Stoop and Meijer (in press-a) proposed to simulate the asymptotic sampling distribution for a given $\hat{\theta}$ through parametric bootstrapping. Given a fixed $\hat{\theta}$-value, P&P and adaptive response vectors were generated and the distribution of the (one-sided) significance probabilities was determined on the basis of $\hat{\theta}$. For P&P tests the results were promising in the sense that the empirical type I errors were in agreement with the nominal type I errors. However, for a CAT the empirical type I errors were too low, which hamper the use of these statistics in a CAT environment.

McLeod and Lewis (1998) examined whether examinees were successful in attaining higher test scores in a CAT, when they had preknowledge of some of the items that were used. Item preknowledge was modeled by a modified three-parameter logistic IRT model, and the probability of a correct response was a combination of the probability of obtaining

a correct response based on preknowledge and the probability of a correct response based on the ability of the examinee. An assumption of the model was, that the probability of a correct response was equal to one when the item was memorized. McLeod and Lewis (1998) proposed to use a posterior log-odds ratio to identify examinees using item preknowledge. However, the effectiveness of this ratio to detect nonfitting item score patterns could not yet be investigated because of the unknown null distribution and thus of the critical values for the log-odds ratio.

Both Van Krimpen-Stoop and Meijer (in press-b) and Bradlow, Weiss, and Cho (1998) proposed to use statistical process control techniques to detect person-misfit in a CAT. Van Krimpen-Stoop and Meijer (in press-b) proposed statistics to investigate person fit in an on-line application and after complete administration of a CAT. Critical values to classify a score pattern as nonfitting were determined by means of a simulation study, which were found to be stable across θ values. Bradlow et al. (1998) discussed several Bayesian methods to simulate a norming distribution that can be used to classify an item score pattern as fitting or nonfitting. They discussed the use of a posterior predictive density and the use of a prior predictive density to generate a distribution for the item score patterns and thus for the person-fit statistics. Because they did not have an item selection algorithm available, they used a permutation distribution to generate a distribution for the item score patterns and illustrated the use of this distribution to detect different types of nonfitting response behavior on empirical CAT data. Although simulation algorithms are useful, they may be computerintensive, and in some situations difficult to apply (for example, in the situation discussed in Bradlow et al., 1998 when the researcher has no disposal to the item selection algorithm). An alternative is to use a theoretical distribution to classify score patterns as fitting or nonfitting. In this paper, we will investigate under which conditions a theoretical distribution can be used when statistics from statistical process control are used in a CAT and which test and person characteristics influence this distribution.

New Person-Fit Statistics

In a CAT, using IRT models of the form of Equation 1, items are often selected for which the probability of correctly answering an item is close to 0.5. As a result, an alternation of

correct and incorrect item scores is expected, and strings of correct or incorrect item scores may indicate nonfitting response behavior. Person-fit statistics are often defined in terms of the residual of the observed and expected item scores (Equation 3), where a correct score (1) results in a positive residual and an incorrect score (0) in a negative residual. For example, suppose an examinee with average θ value takes a test with preknowledge of the items administered in the last part of the CAT. As a result, in the first part of the test the responses will be an alternation of zeros and ones, whereas in the second part more and more items are correctly answered due to item preknowledge. Then for this examinee, runs of positive residuals are expected in the second part of the item score pattern. To detect such score patterns, we will define person-fit statistics that are sensitive to strings of positive or negative residuals of observed and expected item scores, corrected for the use of $\hat{\theta}$. Two different strategies are used to define a person-fit statistic. Person-fit statistics are determined based on (1) the responses to all the items in a CAT, and (2) the responses to a subset of items in a CAT.

General Form of the $Z-$Statistics

Consider an item pool consisting of $i = 1, ..., I$ items and assume that a CAT is administered with fixed test length (for a discussion of the pro's and con's of fixed- and variable-length tests in a CAT, see Davey, Pommerich, & Thompson, 1999). Let X_{i_k} denote the response to item i_k, the kth item administered in the CAT, with $k = 1, ..., N$ with N the final test length. Define

$$W(\theta) = \sum_{k=1}^{N} [X_{i_k} - P_{i_k}(\theta)]$$

as the sum of the residuals of the observed and expected item scores in a CAT. It can be shown (Snijders, in press) that, provided that θ is known,

$$Z(\theta) = \frac{W(\theta)}{\sqrt{\sum_{k=1}^{N} P_{i_k}(\theta) Q_{i_k}(\theta)}} \sim N(0,1) \text{ for } N \to \infty.$$

However, in practice θ is unknown and has to be estimated. Molenaar and Hoijtink (1990) (see also Snijders, in press) showed that when θ is replaced by $\hat{\theta}$, the null distribution of

W and thus also of Z is affected.

Let, for notational simplicity, $P_{i_k} = P_{i_k}(\cdot)$, $Q_{i_k} = Q_{i_k}(\cdot)$, $P'_{i_k} = \partial P_{i_k}/\partial\theta$, and $P''_{i_k} = \partial^2 P_{i_k}/\partial\theta^2$. Consider S subsets of items, each subset consisting of n items, where the first subset of items contains the first n items administered, the second subset contains the next n items administered, and so forth. The number of items, n, in each subset is chosen so that $S = \frac{N}{n}$ is an integer; that is, S complete subsets of n items are administered. For the sake of simplicity, we assume subsets of equal size, but the theoretical results obtained in this paper can easily be extended in terms of subsets of unequal size. Let $s = 1, ..., S$ be the index for the subsets. Define $\hat{\theta}_N$ as the final ability estimate. Furthermore, let $f = N$ indicate that the fit statistic is calculated across all items in the test and $f = s$ that the statistic is calculated across the items in subset s.

Define

$$W_f(\cdot) = \sum_{k=1}^{N} [X_{i_k} - P_{i_k}]\, w_{i_k,f},$$

where, for $f = N$

$$w_{i_k,N} = 1, \text{ for all items } k,$$

and for $f = s$

$$w_{i_k,s} = \begin{cases} 1 & \text{for all items } k \text{ in subset } s \\ 0 & \text{otherwise} \end{cases}$$

Snijders (in press) showed that, provided that some regularity conditions hold (that is, $\exists M < \infty$ such that $\frac{1}{M} < a_i < M$ and $|b_i| < M$ for all i),

$$Z_f(\cdot) = \frac{W_f(\cdot) + c_f(\cdot)\, r_0(\cdot)}{\sqrt{\sum_{k=1}^{N} \tilde{w}^2_{i_k,f}(\cdot)\, P_{i_k} Q_{i_k}}} \sim N(0,1) \text{ for } N \to \infty, \tag{4}$$

where, for weighted maximum likelihood estimator (Warm, 1989), $\hat{\theta}_N$, and the 2-PLM

$$r_0(\cdot) = \frac{\sum_{k=1}^{N} P'_{i_k} P''_{i_k}/P_{i_k}Q_{i_k}}{\sum_{k=1}^{N} \left(P'_{i_k}\right)^2/P_{i_k}Q_{i_k}} = \frac{\sum_{k=1}^{N} a^3_{i_k} P_{i_k} Q_{i_k} (Q_{i_k} - P_{i_k})}{2\sum_{k=1}^{N} a^2_{i_k} P_{i_k} Q_{i_k}},$$

$$
\begin{aligned}
r_{i_k}(\cdot) &= \frac{P'_{i_k}}{P_{i_k}Q_{i_k}} = a_{i_k}, \\
c_f(\cdot) &= \frac{\sum_{k=1}^{N} P'_{i_k} w_{i_k,f}}{\sum_{k=1}^{N} P'_{i_k} r_{i_k}} = \frac{\sum_{k=1}^{N} a_{i_k} P_{i_k} Q_{i_k} w_{i_k,f}}{\sum_{k=1}^{N} a_{i_k}^2 P_{i_k} Q_{i_k}}, \text{ and} \\
\tilde{w}_{i_k,f}(\cdot) &= w_{i_k,f} - c_f(\cdot) r_{i_k}(\cdot) = w_{i_k,f} - a_{i_k} c_f(\cdot).
\end{aligned}
$$

Two Strategies

In the first strategy, the statistic $Z_N\left(\hat{\theta}_N\right)$ is determined; this statistic uses all administered items and the final ability estimate $\hat{\theta}_N$. Then, provided that N is large enough, an examinee can be classified as nonfitting at, for example, significance level $\alpha = 0.01$ when $Z_N > |2.58|$, where 2.58 is the critical value of the standard normal distribution with two-sided $\alpha = 0.01$. A disadvantage of Z_N is that a run of negative (positive) residuals may be compensated for by a run of positive (negative) residuals.

Therefore, to minimize this effect, in the second strategy the item score pattern is divided into disjoint subsets consisting of $n > 1$ items. For each subset $Z_s\left(\hat{\theta}_N\right)$ is determined, and provided that N and n are large enough, critical values can be based on the standard normal distribution. For example, a CAT of $N = 30$ items can be divided into three subsets each consisting of 10 items, where the first subset contains the first 10 administered items, the second subset contains items $11 - 20$, and the third subset contains items $21 - 30$. Then, Z_s can be computed for each subset using $\hat{\theta}_N$ and it can be investigated whether nonfitting behavior occurred on the first 10 items due to, for example, warming-up or item preview. When warming-up occurred, in the first part of the CAT a number of items will be answered incorrectly, resulting in a negative value of the statistic. Thus, nonfitting response behavior is checked for each subset.

By investigating whether the individual scores of Z_s fall in between two bounds, as described in the second strategy, it is investigated whether a shift or change occurred in the mean of Z_s compared with the assumed mean (i.e., zero) of the standard normal distribution. This shift has to be of considerable size to let the value of Z_s fall outside the bounds. As a result, this procedure will be insensitive to small shifts in the mean. Moreover, a disadvantage of using Z_s is that because only part of the CAT is considered the

power to detect nonfitting score patterns may be low. An alternative is to use a cumulative sum procedure where the values of Z_s are accumulated in a clever way, and all item scores in the test are taken into account.

Cumulative Sum Procedure

A procedure from Statistical Process Control which accumulates (standard normally distributed) statistics is the cumulative sum procedure (CUSUM), originally proposed by Page (1954). An assumption underlying the CUSUM procedure is, that the process is assumed to be in statistical control; that is, the variable or statistic being measured has a stable distribution, for example, a normal distribution with a stable mean and variance. In the CUSUM procedure, sums are accumulated as follows. Let Z_t be the value of statistic Z obtained from a sample of size n at time point t, where Z_t is assumed to be independently identically distributed. Let R be a predefined reference value. Then, the two-sided CUSUM procedure can be written in terms of C_t^+ and C_t^-, where

$$\begin{aligned} C_t^+ &= \max\left[0, (Z_t - R) + C_{t-1}^+\right] \text{ and} \\ C_t^- &= \min\left[0, (Z_t + R) + C_{t-1}^-\right], \end{aligned}$$

with starting values $C_0^+ = C_0^- = 0$. Thus, the CUSUM procedure starts as soon as $|Z_t| > R$. Note that the cumulations can be running on both sides concurrently, where the sum of positive values of $(Z_t - R)$ is reflected by C_t^+ and the sum of negative values of $(Z_t + R)$ by C_t^-. Let h denote some threshold value. The process is 'out of control' when $C^+ > h$ or $C^- < -h$ at some time point and 'in control' otherwise; that is, after a run of positive or a run of negative values of the statistic, the process becomes 'out of control'.

The values of R and h are often based on the assumption that the Z_t-values are independently (asymptotically) standard normally distributed. The value of R is usually selected as one-half of the magnitude of the shift (in Z_t-units) one wishes to detect; for example, $R = 0.5$ is the appropriate choice for detecting a shift of one times the standard deviation of Z_t. Based on the chosen values of significance level α and reference value R, boundaries h and $-h$ of the CUSUM procedure can be determined. Note that, although the standard normally distributed statistic, Z, is used, C^+ and C^- are not standard normally

distributed. The critical value h for the two-sided CUSUM procedure can be determined by solving Siegmund's approximation (Siegmund, 1985)

$$\frac{2}{\alpha} = \frac{\exp\left[2R\left(h + 1.166\right)\right] - 2R\left(h + 1.166\right) - 1}{2R^2}$$

where R is the (fixed) reference value and α is the (fixed) two-sided significance level. In Table 1 values of h are given for $R = 0.5$ and 1.0 for α's between .05 and .01.

Insert Table 1 about here

For an appropriate person-fit statistic Z, a CUSUM procedure is sensitive to runs of positive or negative values of that statistic and becomes 'out of control' when a number of positive or negative values of the statistic occurs. When in the CUSUM procedure the statistic $Z_s\left(\hat{\theta}_N\right)$ is used, that is, the standardized residual of observed and expected scores of the items in subset s, an examinee may be classified as nonfitting when a number of consecutive positive or negative values of $Z_s\left(\hat{\theta}_N\right)$ occurs. Using $Z_s\left(\hat{\theta}_N\right)$ results in a CUSUM procedure that can be applied after complete administration of the CAT and can be written as

$$C_s^+ = \max\left[0, \left(Z_s\left(\hat{\theta}_N\right) - R\right) + C_{s-1}^+\right] \tag{5}$$

$$C_s^- = \min\left[0, \left(Z_s\left(\hat{\theta}_N\right) + R\right) + C_{s-1}^-\right] \quad . \tag{6}$$

with starting values $C_0^+ = C_0^- = 0$. An examinee can be classified as nonfitting the IRT model when $C_s^+ > h$ or when $C_s^- < -h$ after some subset s.

Simulation Studies

This study is divided into three parts. First, the nominal and empirical type I errors are compared of the statistics $Z_N\left(\hat{\theta}_N\right)$, $Z_s(\hat{\theta}_N,)$, and the CUSUM procedure. Second, it is investigated whether Z_N and the CUSUM procedure are confounded with θ. This is important because person-fit studies (e.g., Molenaar and Hoijtink, 1990, Reise, 1995) have shown that statistics are confounded with θ and as a result examinees are classified as nonfitting depending on θ which is obviously undesirable. Finally, the detection rates are investigated for several types of nonfitting score patterns.

In all simulation studies, an item pool of 400 items fitting the 2-PLM with $a_i \sim N(1; 0.2)$ (truncated at the interval $\langle 0; 3]$) and $b_i \sim U(-3.5; 3.5)$ was used to simulate the adaptive item score patterns. True θ was assumed to be standard normally distributed (truncated at the interval $[-3; 3]$).

A fitting item score pattern was simulated as follows. First, true θ of a simulee was drawn from a standard normal distribution. Then, the first item of the CAT selected was the item with maximum information given $\theta = 0$. For this item, $P(\theta)$, according to Equation 1 was determined. To simulate the answer (1 or 0), a random number y from the uniform distribution on the interval $[0, 1]$ was drawn; when $y < P(\theta)$ the response to the item was subset to 1 (correct response), 0 otherwise. The first four items of the CAT were selected with maximum information for $\theta = 0$, and based on the responses to these four items, $\hat{\theta}$ was obtained using weighted maximum likelihood estimation (Warm, 1989). The next items were selected to have maximum information at $\hat{\theta}_{i_k}$. For that item, $P(\theta)$ was computed, a response was simulated, and $\hat{\theta}_{i_k}$ was updated. This procedure was repeated until a test of N items was obtained.

Study 1: Empirical vs Nominal Type I Errors

Method

$Z_N\left(\hat{\theta}_N\right)$

Datasets containing 500 fitting adaptive item score patterns were constructed each with a fixed test length N, with $N = 10, 20, 30, 40,$ and 50. For each item score pattern, statistic $Z_N\left(\hat{\theta}_N\right)$ was computed, resulting in 500 values of Z_N for each dataset. For each dataset, the empirical type I error was determined as the percentage of item score patterns that attained a value of Z_N larger than $|2.58|$; that is, the critical value of the standard normal distribution at significance level $\alpha = 0.01$.

$Z_s\left(\hat{\theta}_N\right)$ *and CUSUM*

A dataset containing 500 fitting adaptive item score patterns was constructed with a fixed test length $N = 40$ and the size of the subsets was set to $n = 20, 10,$ or 5. For each item score pattern, statistic Z_s was determined for each subset s. The empirical type I error for subset s was determined as the percentage of item score patterns that attained a value of Z_s larger than $|2.58|$. Also for each item score pattern the CUSUM procedure was performed

using statistic Z_s and $R = 0.5$. The empirical type I error for the CUSUM procedure was determined as the percentage of item score patterns for which $C^+ > 3.49$ or $C^- < -3.49$ at some subset s; that is, $h = 3.49$ is the critical value of the CUSUM procedure with $R = 0.5$ at significance level $\alpha = 0.01$. For Z_N, Z_s and CUSUM, the procedure was repeated ten times for each dataset, in order to obtain the mean and standard deviation of the type I errors.

Results

$Z_N\left(\hat{\theta}_N\right)$

In Table 2 the mean and standard deviation of the empirical type I errors for $Z_N\left(\hat{\theta}_N\right)$ are given at a nominal level of $\alpha = 0.01$. Table 2 shows that the empirical type I errors were in agreement with the nominal type I errors for tests with length 40 and 50, and for tests of 10, 20, and 30 items they were a little smaller than the nominal error rates.

$Z_s\left(\hat{\theta}_N\right)$ *and CUSUM*

In Table 3 the mean and standard deviation of the empirical type I errors for Z_s and the CUSUM procedure are given at $\alpha = 0.01$. Table 3 shows that the empirical type I errors of Z_s for $s = 1$ and for all n are a little smaller than the nominal type I error. It also shows that for $s > 1$ the empirical type I errors are considerably smaller than expected. For the CUSUM, for $n = 10$ the empirical type I error is similar to the nominal type I error, whereas for $n = 20$ it is smaller and for $n = 5$ it is larger. The smaller type I errors are due to the fact that the CUSUM procedure is based on only a few subsets which might result in low power. Some additional simulations, using more subsets, illustrated this effect. For example, increasing test length to 80 items with subsets of size $n = 20$ a mean type I error rate of 0.010 with standard deviation 0.04 was found. On the other hand, for $n = 5$, the subsets are too small to guarantee an accurate approximation of Z_s to the standard normal distribution. Using the CUSUM procedure it is thus important to use (1) subtests that are not too small and (2) an adequate number of measurement points.

Insert Tables 2 and 3 about here

Study 2: Dependence on θ

Method

Five datasets containing 500 fitting score patterns were constructed with fixed test length $N = 40$, where for each dataset θ was set to $-2, -1, 0, 1$, or 2. For each score pattern, statistic $Z_N\left(\hat{\theta}_N\right)$ was determined, and the CUSUM procedure was performed with subsets size $n = 10$ and $R = 0.5$. For each dataset, the empirical type I error for Z_N was determined as the percentage of item score patterns that attained a value of Z_N larger than $|2.58|$ and for the CUSUM procedure as the percentage of item score patterns for which $C^+ > 3.49$ or $C^- < -3.49$.

Results

In Table 4 the mean and standard deviation of the empirical type I errors at nominal level $\alpha = 0.01$ for Z_N and CUSUM, conditional on θ, are given. Table 4 shows that, across θ, the empirical type I errors for Z_N are similar to the nominal type I error. For the CUSUM procedure, the empirical type I errors are quite comparable across θ, although there are small deviancies between the empirical and nominal type I errors for $\theta = \pm 2$.

Insert Table 4 about here

Study 3: Detection Rates

Method

Datasets containing 500 nonfitting item score patterns were constructed for a test of $N = 40$. Nonfitting score patterns were simulated for a two-dimensional value of θ. It was assumed that during the first half of the test an examinee had another θ value than during the second half. Datasets containing score patterns with a two-dimensional θ were simulated by drawing the first ability value, θ_1, from the standard normal distribution truncated $[-3; 3]$, and the second ability value was determined as $\theta_2 = \theta_1 + r$, where $r = -2, -1.5, -1, 1, 1.5$, or 2 were used. Thus, during the first half of the test $P(\theta_1)$ was used and during the second half $P(\theta_2)$ was used to simulate the item scores to the items. Examples of response behavior with $r > 0$ are item preknowledge of items in the second half of the CAT or warming-up in the first half of the test. Examples of $r < 0$ are

preknowledge of the items in the first half of the CAT or carelessness in the last part of the CAT.

For each item score pattern in each dataset, Z_N, Z_s, and the CUSUM procedure with $n = 10$ and $R = 0.5$ were computed. Z_s was used with $s = 1$ and $s = 2$, because in Study 2 it was shown that using $s = 3, 4$ resulted in type I errors that were much too small. An item score pattern was classified as nonfitting at significance level $\alpha = 0.01$, when $Z_f > |2.58|$ where $f = N$ or s, or when $C^+ > 3.49$ or $C^- < -3.49$ at some subset s. For each dataset, the detection rate was determined as the percentage of item score patterns classified as nonfitting. For each value of r, 10 datasets containing 500 nonfitting item score patterns were generated and the mean and standard error were computed. The same procedure was performed using 10 datasets containing 500 fitting adaptive response vectors ($r = 0$).

Results

In Table 5 the detection rates of $Z_N\left(\hat{\theta}_N\right)$, $Z_s\left(\hat{\theta}_N\right)$ and the CUSUM procedure are given. Table 5 shows that for fitting item score patterns ($r = 0$), the empirical type I error is around 0.01 as expected at level $\alpha = 0.01$. The detection rates for the CUSUM procedure are considerably higher than the detection rates for Z_N and Z_s for all s, for all values of $r \neq 0$. Interesting is, that the detection rates for Z_s at set $s = 2$ are higher than Z_N and Z_s with $s = 1$ for all values of $r \neq 0$ except $r = -1$. This can be explained by the observation that $\hat{\theta}$ at the beginning of the test is not yet estimated accurately. Consequently, the residuals in Equation 3 are small and Z_s will not obtain large values.

Insert Table 5 about here

Discussion

In this study, we examined the application of person-fit statistics using known statistical distributions in the context of a CAT. Two strategies were investigated: person-fit statistics were determined based on (1) the responses to all the items (statistic Z_N and the CUSUM procedure), and (2) the responses to a subset of items (statistic Z_s).

Study 1 showed that the critical values of the standard normal distribution can be used

for Z_N, although for small test length ($N = 10, 20$) empirical type I errors are somewhat smaller than expected. Results also showed that for $N = 40$ and $n = 10$ the critical values (Table 1) can be used for the CUSUM procedure. However, care should be taken when the size of the subsets n is large compared with N. At a fixed test length, increasing the number of items in a subset (increasing n), will result in a smaller number of subsets S. For a small number of subsets, the CUSUM procedure is based on only a few subsets which might result in low power. On the other hand, when n is small, the number of subsets increases and the asymptotic normality of the statistic might no longer hold. This combined with large N results in an accumulation of error and in an inaccurate empirical type I error of the CUSUM procedure. An assumption of the CUSUM procedure is that the statistics Z_s are independently distributed. Here, the final ability estimate $\hat{\theta}_N$ is based on all item scores, and $\hat{\theta}_N$ is used for computing each value of Z_s. As a result, some dependence might be introduced. This may explain the somewhat inflated empirical type I error rates for large N in combination with small n.

Based on the simulation studies reported here and some additional experience with the procedure, we propose the following rule of thumb: the CUSUM procedure can be used when $n \geq 10$ and $S = 3, 4$ or 5. The best strategy is to construct subsets as large as possible and to divide the test into 3, 4, or 5 subsets. This protects the user against incorrect type I errors that will result in liberal or conservative classification of nonfitting item score patterns.

In this paper we examined the detection rate for nonfitting item score patterns with a two-dimensional θ-value and found that the proposed statistics, especially the CUSUM procedure, were sensitive to this type of nonfitting response behavior; that is, the detection rates were between 0.12 and 0.67 at $\alpha = 0.01$ for the CUSUM.

All person-fit statistics proposed in this paper need to be determined after complete administration of the CAT due to the use of the final ability estimate $\hat{\theta}_N$. An alternative is to construct similar statistics where the updated ability estimate is used; this results in an on-line application of person fit. However, this alternative is less attractive than the assessment of person fit after complete administration of a CAT for at least two reasons. First, the best estimate of true θ available is the final ability estimate $\hat{\theta}_N$. Additional simulation studies showed that using updated ability estimates resulted in large differences

between empirical and nominal type I error rates. Second, the fit of an item score pattern given the final ability estimate will be used most often in practice because decisions in job and educational selection using personality tests and achievement tests are based on the final estimate.

The statistics proposed in this study are sensitive to runs of correct or incorrect item scores. An alternative to the proposed statistics might be to use nonparametric statistics such as the number of runs or the length of the longest runs in a series of alternatives; a crucial assumption underlying the distribution of these statistics, however, is that the expected item score is equal to 0.5 for all items in the test. Especially in the beginning of a CAT, the expected item score may deviate from 0.5. Our experience is that this seriously influences the error rates. The statistics proposed in this study are less sensitive to violation of this assumption, because here the residuals of the observed and expected score on the basis of an IRT model are taken into account.

Future research may also investigate subgroupings of items, for example subgroupings on the basis of content area. Doing so, person-fit statistics can be used as diagnostic tools. In practice two research strategies may be followed depending on the application envisaged. In the first situation the researcher may like to have information about the answering behavior on specific subsets of items. When analyzing the score patterns on, for example, an achievement test a researcher may only be interested in detecting unusual score patterns on particular subsets of items because these patterns may be indicative of a particular kind of nonfitting behavior. Examples may be test preview and lack of concentration. In this case, the researcher has a priori expectations with respect to score patterns and he/she can test these hypotheses using the theoretical distributions discussed in this paper. A limitation may be that the subsets may not be too small, say not smaller than 10-15 items. In the second situation, the researcher may have no idea about which type of aberrance he/she may expect to find. In that case, first a regular overall fit statistic such as Z_N or the CUSUM procedure can be used. Next these nonfitting patterns can further be investigated using statistics that are sensitive to nonfitting items scores in subsets of the test.

Author Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the authors and not necessarily reflect the position or policy of LSAC.

References

Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. Dekker, New York.

Bradlow, E. T., Weiss, R. E., and Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association*, 93, 910–919.

Davey, T., Pommerich, M., and Thompson, T. D. (1999). *Pretesting alongside an operational CAT*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Montreal, Canada.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67–86.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: principles and applications*. Kluwer-Nijhoff, Boston.

Levine, M. V. and Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161–176.

Levine, M. V. and Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269–290.

McLeod, L. D. and Lewis, C. (1998). *A Bayesian approach to detection of item preknowledge in a CAT*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Diego, CA.

Meijer, R. R., Molenaar, I. W., and Sijtsma, K. (1994). Item, test, person and group characteristics and their influence on nonparametric appropriateness measurement. *Applied Psychological Measurement*, 18, 111–120.

Molenaar, I. W. and Hoijtink, H. (1990). The many null distributions of person fit

indices. *Psychometrika*, 55, 75–106.

Molenaar, I. W. and Hoijtink, H. (1996). Person fit and the Rasch model, with an application of knowledge of logical quantors. *Applied Measurement in Education*, 9, 27–45.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, 21, 115–127.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41, 100–115.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, 19, 213–229.

Reise, S. P. and Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, 15, 217–226.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.

Snijders, T. (in press). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika*.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (in press-a). Simulating the null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (in press-b). Detecting person misfit in adaptive testing using statistical process control techniques. In van der Linden, W. J. and Glas, C. A. W., (Eds.), *Computerized adaptive testing: Theory and practice*. Kluwer-Nijhoff, Boston, MA.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54.

Wright, B. D. and Stone, M. H. (1979). *Best test design. Rasch measurement*. Mesa Press, Chicago.

Table 1

Critical values h for several values of R and α

	h	
α	R = 0.5	R = 1.0
.0500	2.02	1.06
.0250	2.64	1.39
.0100	3.49	1.84
.0050	4.16	2.18
.0025	4.84	2.53
.0010	5.75	2.98

Table 2

Mean and standard deviation of the empirical type I errors at $\alpha = 0.01$ for $Z_N\left(\hat{\theta}_N\right)$

	Z_N	
N	**mean**	**SD**
10	.005	.008
20	.006	.008
30	.005	.009
40	.010	.004
50	.009	.004

Table 3

Mean and standard deviation of the empirical type I errors at $\alpha = 0.01$ for $Z_s\left(\hat{\theta}_N\right)$ and the CUSUM procedure for a CAT with $N = 40$

n		s	mean	SD
20	Z_s	1	.006	.004
		2	$< .001$	
	CUSUM		$< .001$	
10	Z_s	1	.005	.003
		2	.002	.001
		3	$< .001$	
		4	$< .001$	
	CUSUM		.008	.004
5	Z_s	1	.004	.003
		2	.002	.002
		3	.001	.001
		4	$< .001$	
		5	$< .001$	
		6	$< .001$	
		7	$< .001$	
		8	$< .001$	
	CUSUM		.067	.015

Table 4

Mean and standard deviation of the empirical type I errors at $\alpha = 0.01$ for $Z_N\left(\hat{\theta}_N\right)$ and CUSUM procedure, conditional on θ, for a CAT with $N = 40$ and $n = 10$

		Z_N		CUSUM	
		mean	**SD**	**mean**	**SD**
$\theta =$	-2	.011	.003	.002	.002
	-1	.009	.004	.007	.003
	0	.009	.005	.008	.003
	1	.008	.003	.005	.003
	2	.010	.004	.002	.001

Table 5

Mean and standard deviation (in brackets) of the detection rates of $Z_N\left(\hat{\theta}_N\right)$, $Z_s\left(\hat{\theta}_N\right)$ and the CUSUM procedure for a CAT with $N = 40$ and $n = 10$

	r						
	−2	**−1.5**	**−1**	**0**	**1**	**1.5**	**2**
$\mathbf{Z}_N$	.163	.108	.053	.012	.042	.079	.122
	(.018)	(.011)	(.012)	(.004)	(.010)	(.020)	(.015)
CUSUM	.667	.360	.123	.006	.116	.348	.641
	(.021)	(.024)	(.016)	(.005)	(.020)	(.028)	(.018)
$\mathbf{Z}_s, \mathbf{s = 1}$	.194	.092	.033	.003	.031	.094	.214
	(.015)	(.014)	(.008)	(.003)	(.007)	(.011)	(.019)
$\mathbf{Z}_s, \mathbf{s = 2}$	.552	.229	.051	.001	.056	.213	.494
	(.023)	(.022)	(.011)	(.001)	(.007)	(.020)	(.020)

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

BEST COPY AVAILABLE